

AD-A208 77M

THIS FILE COPY

(4)

CRM 88-236 / December 1988

## RESEARCH MEMORANDUM

# A MAXIMUM-LIKELIHOOD PROCEDURE FOR DEVELOPING A COMMON METRIC IN ITEM-RESPONSE THEORY

D. R. Divgi

A Division of

**CNA**

Hudson Institute

**CENTER FOR NAVAL ANALYSES**

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22301-0268

This document has been approved  
for public release and sale in  
unlimited quantities.

89 6 02 026

DTIC  
ELECTE  
JUN 02 1989  
S E

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION  
UNLIMITED

*Work conducted under contract N00014-87-C-0001.*

This Research Memorandum represents the best opinion of CNA at the time of issue.  
It does not necessarily represent the opinion of the Department of the Navy.

REPORT DOCUMENTATION PAGE				
1a. REPORT SECURITY CLASSIFICATION <b>UNCLASSIFIED</b>		1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for Public Release; Distribution unlimited		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE		5. MONITORING ORGANIZATION REPORT NUMBER(S)		
4. PERFORMING ORGANIZATION REPORT NUMBER(S)  CRM 88-236		7a. NAME OF MONITORING ORGANIZATION  Commanding General, Marine Corps Combat Development Command		
6a. NAME OF PERFORMING ORGANIZATION  Center for Naval Analyses		6b. OFFICE SYMBOL (If applicable) CNA		7b. ADDRESS (City, State, and ZIP Code) Warfighting Center Quantico, Virginia 22134
6c. ADDRESS (City, State, and ZIP Code)  4401 Ford Avenue Alexandria, Virginia 22302-0268		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER  N00014-87-C-0001		
8a. NAME OF FUNDING ORGANIZATION  Office of Naval Research		8b. OFFICE SYMBOL (If applicable) ONR		10. SOURCE OF FUNDING NUMBERS
8c. ADDRESS (City, State, and ZIP Code)  800 North Quincy Street Arlington, Virginia 22217		PROGRAM ELEMENT NO. 65153M	PROJECT NO. C0031	TASK NO. .
11. TITLE (Include Security Classification) A Maximum-Likelihood Procedure for Developing a Common Metric in Item-Response Theory				
12. PERSONAL AUTHOR(S) D.R. Divgi				
13a. TYPE OF REPORT  Final		13b. TIME COVERED FROM TO		14. DATE OF REPORT (Year, Month, Day) December 1988
15. PAGE COUNT  12				
16. SUPPLEMENTARY NOTATION				
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD	GROUP	SUB-GROUP	ASVAB (Armed Services Vocational Aptitude Battery), CAT (Computerized Adaptive Testing), Maximum likelihood estimation, Parameters, Statistical analysis, Statistical processes, Test methods	
12	03			
05	08			
19. ABSTRACT (Continue on reverse if necessary and identify by block number)  Because the ability scale in item-response theory is arbitrary, if two item pools are calibrated in two different samples, their parameter estimates must be placed on a common metric using items administered in both calibrations. In this memorandum, a maximum-likelihood procedure for doing so is derived and illustrated.				
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input type="checkbox"/> UNCLASSIFIED / UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION <b>UNCLASSIFIED</b>	
22a. NAME OF RESPONSIBLE INDIVIDUAL Colonel Preston			22b. TELEPHONE (Include Area Code)  22c. OFFICE SYMBOL MCCDC	

**CNA****CENTER FOR NAVAL ANALYSES**

A Division of Hudson Institute

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

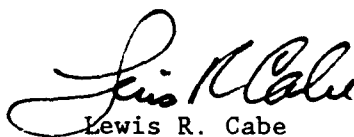
1 February 1989

## MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 88-236

Encl: (1) CNA Research Memorandum 88-236, *A Maximum-Likelihood Procedure for Developing a Common Metric in Item-Response Theory*, by D. R. Divgi, Dec 1988

1. Enclosure (1) is forwarded as a matter of possible interest.
2. A computerized adaptive version of the Armed Services Vocational Aptitude Battery has been developed. This development required application of item-response theory to two different item pools administered to non-equivalent samples. In such cases, the ability scales in the two samples must be placed on a common metric. A maximum-likelihood procedure for doing so is presented and illustrated with examples.

Lewis R. Cabe  
Director  
Manpower and Training ProgramDistribution List:  
Reverse page

Accession For	
IEEE GNA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/ _____	
Availability Codes	
Dist	Avail and/or Special
<b>A-1</b>	

Subj: Center for Naval Analyses Research Memorandum 88-236

Distribution List

SNDL

A1	ASSTSECNAV MRA
A1	DASN MANPOWER (2 copies)
A2A	CNR
A6	HQMC MPR
	Attn: M
	Attn: MP
	Attn: MR
	Attn: MA (2 copies)
	Attn: MPP-39
A6	HQMCRA
A6	HQMC AVN
A6	CG MCRDAC, Washington
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL (2 copies)
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FKQ6D	NAVPERSRANDCEN
	Attn: Technical Director (Code 01)
	Attn: Director, Testing Systems (Code 63)
	Attn: Technical Library
	Attn: Director, Personnel Systems (Code 62)
	Attn: CAT/ASVAB PMO
	Attn: Manpower Systems (Code 61)
FT1	CNET
V12	CG MCRDAC, Quantico
	Attn: Director, Development Center Plans Division (Code D08)
	(2 copies)
	Attn: Commanding General
V12	CGMCCDC
	Attn: Training and Education Center
OPNAV	
OP-01	
OP-11	
OP-13	
OP-15	

CRM 83-236 / December 1988

# **A MAXIMUM-LIKELIHOOD PROCEDURE FOR DEVELOPING A COMMON METRIC IN ITEM-RESPONSE THEORY**

D. R. Divgi

*A Division of*



*Hudson Institute*

---

**CENTER FOR NAVAL ANALYSES**

*4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268*

#### ABSTRACT

Because the ability scale in item-response theory is arbitrary, if two item pools are calibrated in two different samples, their parameter estimates must be placed on a common metric using items administered in both calibrations. In this memorandum, a maximum-likelihood procedure for doing so is derived and illustrated.

## EXECUTIVE SUMMARY

The Department of Defense has developed a computerized adaptive testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) in the Accelerated CAT-ASVAB Project (ACAP). Use of the CAT requires a large pool of items for each subtest. For Arithmetic Reasoning and Word Knowledge, it became necessary to supplement the original ACAP pool with items from the experimental CAT-ASVAB system developed earlier. This memorandum presents a maximum-likelihood procedure for performing some calculations needed to merge the two item pools.

CAT-ASVAB uses the three-parameter logistic model of item-response theory (IRT). In this model, each person is characterized by an ability parameter  $\theta$  and each test item by three parameters  $a$ ,  $b$ , and  $c$ . The quantities  $a$ ,  $b$ , and  $c$  are called the discrimination, difficulty, and guessing parameters of the item.

The metric of the  $\theta$  scale is arbitrary. One can transform  $\theta$ ,  $a$ , and  $b$  simultaneously in such a way that the probability of answering an item correctly remains unchanged for all persons and items. This creates a practical problem. Suppose two tests are calibrated--that is, their item parameters are estimated, using different samples of examinees. One set of item parameters must be transformed to the metric of the other before the two sets of estimates can be used together. This requires that the tests have some items in common.

Currently available procedures for determining a transformation define a criterion function and minimize it to estimate the transformation parameters. Although reasonable, the criterion function is not based on any principle or related to the larger problem of estimating item parameters.

Item parameters are usually estimated by the method of maximum likelihood. The same approach can be extended to transform the metric of one calibration to that of another. The method is illustrated in this memorandum using four forms each of five ASVAB subtests, which were included in calibrations of both the experimental and ACAP item pools. Results using this method are found to be close to those of an earlier method devised by Stocking and Lord.

Maximum likelihood is a viable procedure that can be used with item pools for future versions of CAT-ASVAB. It requires less computation than the Stocking-Lord method and makes use of information about standard errors of parameter estimates.



## TABLE OF CONTENTS

	Page
Introduction .....	1
Metric Transformation in IRT .....	1
Maximum Likelihood Approach .....	2
Illustration .....	4
References .....	7

## INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) is used to select and classify enlisted personnel. It contains ten subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). The Verbal (VE) subtest is defined as the sum of WK and PC.

The Department of Defense has developed a computerized adaptive testing (CAT) version of the ASVAB in the Accelerated CAT-ASVAB Project (ACAP). Use of the CAT requires a large pool of items for each subtest. For Arithmetic Reasoning and Word Knowledge, it became necessary to supplement the original ACAP pool with items from the experimental CAT-ASVAB system developed earlier [1]. The purpose of this memorandum is to present a maximum-likelihood procedure for performing some calculations needed to merge the two item pools.

## METRIC TRANSFORMATION IN IRT

CAT-ASVAB uses the three parameter logistic model of item-response theory (IRT). In this model, each person is characterized by an ability parameter  $\theta$  and each test item by three parameters  $a$ ,  $b$ , and  $c$ . The probability that a person of ability  $\theta$  will answer an item correctly is given by

$$P(\theta) = c + (1 - c) / [1 + \exp\{1.7a(b - \theta)\}] .$$

The quantities  $a$ ,  $b$ , and  $c$  are called, respectively, the *discrimination*, *difficulty*, and *guessing* parameters of the item.

The metric of the  $\theta$  scale is arbitrary. It is possible to make a linear transformation of  $\theta$ ,  $a$ , and  $b$  in such a way that  $P(\theta)$  remains unchanged. Suppose two tests are calibrated--that is, their item parameters are estimated, using samples of examinees from different populations. One set of item parameters must be transformed to the metric of the other before a useful analysis (e.g., equating) can be performed. This requires that the tests have at least two items in common.

Let estimates from the second calibration be transformed to the metric of the first. Transformed estimates of discrimination ( $a$ ) and difficulty ( $b$ ) parameters are given for each item by

$$a_2^* = a_2 / A , \tag{1}$$

$$b_2^* = Ab_2 + B , \tag{2}$$

and for each person by

$$\theta_2^* = A\theta_2 + B, \quad (3)$$

where \* indicates a transformed value and the subscript refers to the calibration. It is easy to verify that the probability  $P(\theta)$  is invariant under such transformations.

Recent procedures for estimating  $A$  and  $B$  are found in Stocking and Lord [2] and Divgi [3]. These methods estimate the parameters by minimizing a criterion function, which is a weighted sum of squares. The procedures are ad hoc in that the criterion functions, although reasonable, are not based on any principle. The purpose of this memorandum is to relate the estimation of  $A$  and  $B$  to the larger problem of estimating item parameters. This leads to a procedure that, like parameter estimation, is based on the principle of maximizing a likelihood function.

#### THE MAXIMUM-LIKELIHOOD APPROACH

No metric transformation would be necessary if a single joint calibration were performed using both samples at once. The two calibrations provide independent sets of parameter estimates for each item. If one tries to combine them so as to approximate the single set of estimates that a joint calibration would yield, a procedure for metric transformation emerges.

Ideally, all three parameters should be included in the calculations. However, the guessing parameter  $c$  is often difficult to estimate with the sample sizes available in practice. Wainer and Thissen [4] have shown that theoretical standard errors of the estimates of  $c$  can be very high for easy items. For this reason, compromises have to be made: data on different items must be pooled or Bayesian prior distributions must be used to keep the estimates reasonable. Standard errors of these estimates are much smaller than their theoretical values. Hence, given that the  $c$  parameter is not estimated by pure maximum likelihood in the original calibrations, no direct use of it is made in the theory given below.

Let vectors

$$p_1 = (a_1 \ b_1)'$$

and

$$p_2^* = (a_2^* \ b_2^*)'$$

denote the two pairs of estimates for an item common to both tests. They maximize log likelihoods  $L_1$  and  $L_2$  in the two samples. Now suppose a joint calibration is performed and the results transformed to the metric of calibration 1. Denote these estimates by

$$p = (a \ b)' ,$$

which maximize  $L_1 + L_2$ . Therefore  $p$  can be calculated approximately from  $p_1$  and  $p_2^*$ .

If the samples are large, estimates of item parameters are close to their true values. Therefore, if transformation parameters  $A$  and  $B$  are chosen properly,  $p_1$  and  $p_2^*$  are almost equal. In their neighborhood, the log likelihoods of responses observed in the samples are quadratic functions of the parameters. Denote the information matrices, i.e.,  $2 \times 2$  matrices of second derivatives of log likelihood, by  $I_1$  and  $I_2^*$ . (Formulas for computing them in the three-parameter logistic model are given by Lord [5].) Let  $L_{1m}$  and  $L_{2m}$  be maximum values of log likelihoods of responses on the common items in calibrations 1 and 2. For any parameter vector  $p$  near  $p_1$  and  $p_2^*$ , log likelihood  $L_1 + L_2$  for the two samples combined is given by

$$2(L_{1m} + L_{2m} - L_1 - L_2) = \sum [(p - p_1)' I_1 (p - p_1) + (p - p_2^*)' I_2^* (p - p_2^*)] \quad (4)$$

where the sum is taken over all items. Minimizing this quantity over a single item leads to a linear equation for  $p$ . Its solution yields

$$p - p_2^* = (I_1 + I_2^*)^{-1} I_1 (p_1 - p_2^*) .$$

A little matrix manipulation shows that the minimum value of the item's contribution to equation (4) is

$$(p_1 - p_2^*)' S (p_1 - p_2^*) , \quad (5a)$$

where

$$S = I_1 - I_1 (I_1 + I_2^*)^{-1} I_1 . \quad (5b)$$

Multiplication verifies that

$$S = (I_1^{-1} + I_2^{*-1})^{-1} .$$

Thus, for any given  $A$  and  $B$ , after minimizing over  $p$  for each common item,

$$2(L_{1m} + L_{2m} - L_1 - L_2) = \varepsilon (p_1 - p_2^*)' (I_1^{-1} + I_2^{*-1})^{-1} (p_1 - p_2^*) . \quad (6)$$

Minimization of this quantity over  $A$  and  $B$  yields maximum likelihood estimates of the transformation parameters.

The argument leading to expression (6) is strictly correct only if true abilities are known. In practice the maximum likelihood estimates of  $\theta$  are used instead [5], or the likelihood is marginalized by integrating over the distribution of ability (Bock and Aitkin [6]). It does not matter how the likelihood function is calculated; if it yields satisfactory estimates of item parameters, it can be used to compute the information matrices in expression (6).

The criterion function (6) is the same as in Divgi's minimum chi-square method [3]. In addition to supplying a theoretical basis for the minimum chi-square method, the maximum-likelihood approach shows how the guessing parameter  $c$  should be handled. Theoretical information functions involving derivatives with respect to  $c$  often greatly overestimate the true standard errors; hence they are excluded from the theory. Estimates of  $c$  do not appear directly in the criterion function; however, they are used in computing  $2 \times 2$  information matrices for  $a$  and  $b$ .

#### ILLUSTRATION

For each subtest in CAT-ASVAB, the item pool was divided into booklets. Each booklet was administered to a large sample of military

applicants, along with an operational form of the ASVAB. Hence the item calibration provided parameter estimates for operational ASVAB items as well as for the CAT pool. This design was used for the ACAP version of CAT-ASVAB [7] and also for the earlier experimental version [8].

ASVAB forms 9A, 9B, 10A, and 10B were used operationally in both calibrations. Therefore, two sets of parameter estimates are available for each form. Estimates for all subtests in the ACAP calibration and for five subtests in the experimental calibration have been provided to the Center for Naval Analyses by the Navy Personnel Research and Development Center. These five subtests are GS, AR, WK, PC, and MK.

The maximum-likelihood and Stocking-Lord [2] procedures were applied to each form of each of the five subtests. Information matrices needed in the maximum-likelihood method were computed under the assumption that the ability distribution was standard normal in each calibration. The same assumption was made while sampling  $\theta$  values needed in the Stocking-Lord method. The normality assumption is reasonable and used frequently (for example, in the calibration of the ACAP item pool [7]).

The results are presented in table 1. For any given subtest, the results vary little from one form to another and from one method to the other. This is to be expected since all eight values (e.g., for A) are estimates of the same quantity.

The assumptions of the maximum-likelihood approach are reasonable, and its theory is simple. It is only to be expected that its results should agree with the more established Stocking-Lord procedure. The illustration serves primarily as a check on the computer program. It is much harder to decide whether one method is clearly preferable to the other. To do so would require extensive data analyses, which are beyond the scope of this paper. However, as pointed out in [3], the chi-square method involves much simpler computations and, unlike the Stocking-Lord method, makes use of information about the sampling errors of the estimates of item parameters.

TABLE 1

## RESULTS OF MAXIMUM LIKELIHOOD AND STOCKING-LORD PROCEDURES

Subtest	Form	Maximum likelihood		Stocking- Lord	
		A	B	A	B
GS	9A	1.17	-.27	1.14	-.21
GS	9B	1.16	-.28	1.09	-.20
GS	10A	1.09	-.28	1.04	-.19
GS	10B	1.13	-.23	1.19	-.26
AR	9A	1.12	-.30	1.11	-.30
AR	9B	1.17	-.31	1.14	-.31
AR	10A	1.12	-.27	1.13	-.30
AR	10B	1.16	-.35	1.13	-.34
WK	9A	1.14	-.27	1.15	-.30
WK	9B	1.24	-.34	1.22	-.33
WK	10A	1.10	-.30	1.17	-.35
WK	10B	1.16	-.34	1.13	-.33
PC	9A	0.87	-.10	0.99	-.19
PC	9B	1.01	-.26	1.03	-.28
PC	10A	0.96	-.16	1.06	-.25
PC	10B	1.05	-.19	1.11	-.29
MK	9A	1.26	-.45	1.25	-.42
MK	9B	1.32	-.51	1.29	-.45
MK	10A	1.29	-.50	1.25	-.43
MK	10B	1.27	-.45	1.30	-.45

## REFERENCES

- [1] Defense Manpower Data Center, *Minutes of April 1987 Meeting of the Psychometric Committee*, by Bruce Bloxom, 12 May 1987
- [2] M. L. Stocking and F. M. Lord. "Developing a Common Metric in Item Response Theory." *Applied Psychological Measurement* (Spring 1983): 201-210
- [3] D. R. Divgi. "A Minimum Chi-Square Method for Developing a Common Metric in Item Response Theory." *Applied Psychological Measurement* (Dec 1985): 413-415
- [4] Howard Wainer and David Thissen. "Some Standard Errors in Item Response Theory." *Psychometrika*, (Dec 1982): 397-412
- [5] F. M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum, 1980
- [6] R. D. Bock and M. Aitkin. "Marginal Maximum Likelihood Estimation of Item Parameters: An Application of an EM Algorithm." *Psychometrika* (Dec 1981): 443-459
- [7] Air Force Human Resources Laboratory, Armed Services Vocational Aptitude Battery: *Development of an Adaptive Item Pool*, AFHRL-TR-85-19 by J. Stephen Prestwood, C. David Vale, Randy H. Massey and John R. Welsh, Sep 1985
- [8] J. B. Sympson and L. Hartmann. "Item Calibrations for Computerized Adaptive Testing (CAT) Experimental Item Pools," in D. J. Weiss (ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Apr 1985